Köhler, Carmen; Pohl, Steffi; Carstensen, Claus H.

Dealing with item nonresponse in large-scale cognitive assessments. The impact of missing data methods on estimated explanatory relationships

formal und inhaltlich überarbeitete Version der Originalveröffentlichung in:
formally and content revised edition of the original source in:
Journal of educational measurement 54 (2017) 4, S. 397-419, 10.1111/jedm.12154

Dealing with item nonresponse in large-scale cognitive assessments: The impact of missing data methods on estimated explanatory relationships

Carmen Köhler[1], Steffi Pohl[2], and Claus H. Carstensen[3]

[1]German Institute for International Educational Research (DIPF)

[2]Freie Universität Berlin, Germany

[3]Otto-Friedrich-University Bamberg, Germany

Author Note

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within the Priority Programme 1646: Education as a Lifelong Process (Grant No. PO 1655/1-1). The authors thank Katja Buntins, David Kaplan, Ingrid Koller, Irini Moustaki, and the anonymous reviewers for useful advice and discussions.

This paper uses data from the National Educational Panel Study (NEPS) Starting Cohort 6–Adults (Adult Education and Lifelong Learning), doi:10.5157/NEPS:SC6:3.0.1. From 2008 to 2013, the NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research and supported by the Federal States. As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi).

Correspondence concerning this article should be addressed to Carmen Köhler, German Institute for International Educational Research (DIPF), Schloßstraße 29, 60486 Frankfurt, Germany. Phone: +49-(0)69-24708-122. E-Mail: carmen.koehler@dipf.de.

Dealing with item nonresponse in large-scale cognitive assessments – The impact of missing data methods on estimated explanatory relationships

Abstract

Competence data from low-stakes educational large-scale assessment studies allow for evaluating relationships between competencies and other variables. The impact of item-level nonresponse has not been investigated with regard to statistics that determine the size of these relationships (e.g., correlations, regression coefficients). Classical approaches such as ignoring missing values or treating them as incorrect are currently applied in many large-scale studies, while recent model-based approaches that can account for nonignorable nonresponse have been developed. Estimates of item and person parameters have been demonstrated to be biased for classical approaches when missing data are missing not at random (MNAR). In our study, we focus on parameter estimates of the structural model (i.e., the true regression coefficient when regressing competence on an explanatory variable), simulating data according to various missing data mechanisms. We found that model-based approaches and ignoring missing values performed well in retrieving regression coefficients even when we induced missing data that were MNAR. Treating missing values as incorrect responses can lead to substantial bias. We demonstrate the validity of our approach empirically and discuss the relevance of our results.

Keywords: missing data, missing propensity, item response theory, large-scale assessment, simulation study

Educational large-scale assessment studies such as PISA (Programme for International Student Assessment) aim to "provide information on the relative performance of students and on differences between student environments, attitudes, and experiences" (Kastberg, Roey, Lemanski, Chan, & Murray, 2014, p. 1). The outcomes of these assessments can have a major impact on policies and political choices in the educational system. The accurate scaling of educational large-scale assessment tests is therefore of utmost importance. Item nonresponse can pose a threat to the scaling of competences, especially when they relate to the unobserved response—that is, the true value on the item if it had been observed (Mislevy & Wu, 1996). The aim of the current study was to investigate how different treatments of item nonresponse affect relevant outcome measures of educational assessments, such as relationship estimates between competence and an explanatory variable.[1] In the subsequent sections, we introduce the different types of missing item responses, followed by a description of current missing data approaches practiced in large-scale assessments. We further outline a recently developed missing data approach that takes nonignorable missing values into account. We then present findings from previous missing data studies regarding the different missing data approaches; these lead to the research questions and scope of our study.

Not-administered items are planned missing values that result from rotated test designs or from (computer) adaptive testing. Not-administered items due to rotated testing are considered missing completely at random (MCAR), since they depend on neither observed nor on unobserved responses: the test booklets are randomly distributed, and missing values are determined by the test developer (Mislevy & Wu, 1996). In adaptive testing, the test developer defines a selection process for the items based on response sequences on previous items. Missing

[1]Note that this paper refrains from discussing missing values on items from (background) questionnaires and exclusively focuses on item nonresponse in achievement tests.

values due to adaptive testing satisfy the missing at random (MAR) condition, since the probability for an item to be missing depends on the observed responses to previous items but not on unobserved responses (Mislevy & Wu, 1996). According to Rubin (1976), missing values can be ignored in the scaling procedure when MAR (or MCAR) and distinctness between the parameters of the analytical model of interest and the parameters of the model for missingness hold. Accordingly, it is possible to ignore planned missing values due to rotated test designs and adaptive testing.

Compared to planned missing values, unplanned missing values such as *not-reached* and *omitted items* pose a greater challenge to the scaling process. Omitted items are skipped items, which can appear at each section of the test; the term *not-reached item* typically refers to all missing values after the last valid given response (see, e.g., Lord, 1974). Not-reached and omitted items are generally considered missing not at random (MNAR; see, e.g., Glas & Pimentel, 2008; Mislevy & Wu, 1996; Rose, von Davier, & Xu, 2010). Their occurrence is neither determined by the test developer nor the observed responses but might depend on unobserved responses. The assumptions for ignorability are thus violated (Rubin, 1976; Mislevy & Wu, 1996).

Current practices in low-stakes educational large-scale achievement tests involve treating unplanned missing values as incorrect or fractionally correct responses or ignoring them in the scaling (see, e.g., PISA, Adams & Wu, 2002; TIMSS [Third International Mathematics and Science Study], Martin, Gregory, & Stemler, 2000; NAEP [National Assessment of Educational Progress], Allen, Donoghue, & Schoeps, 2001; NEPS [National Educational Panel Study], Pohl & Carstensen, 2012). Research on these types of missing data approaches showed bias on item and person parameter estimates when missing values were scored as incorrect (Culbertson, 2011; DeAyala, Plake, & Impara, 2001; Finch, 2008; Hohensinn & Kubinger, 2011; Holman & Glas, 2005; Pohl et al., 2014; Rose et al., 2010). The method of fractionally correct scoring performed

slightly better but also resulted in bias, especially when missing values were MNAR (DeAyala et al., 2001; Finch, 2008). Item and person parameters were less affected when missing values were simply ignored, but this approach also showed bias when the amount of nonignorable missing values was large (Custer, Sharairi, & Swift, 2012; DeAyala et al., 2001; Holman & Glas, 2005; Rose, 2013; Rose et al., 2010).[2]

Model-based approaches aim to avoid this bias by incorporating a model for the process that causes missing values in the competence measurement model (Holman & Glas, 2005; O'Muircheartaigh & Moustaki, 1999). The approach thus provides the opportunity to address nonignorable nonresponses. The idea is to model an additional manifest or latent variable that describes the occurrence of missing values. This variable is typically referred to as the person's *missing propensity*, which is the examinee's tendency to not reach or omit items (see, e.g., O'Muircheartaigh & Moustaki, 1999). In the latent approach, a two-dimensional item response theory (IRT) model is typically employed. Person ability, θ, is modeled based on the response indicators x_{iv}, where i indexes the items from $i = 1, \ldots, I$, and v indexes the persons from $v = 1, \ldots, V$. The response indicators are defined as

$$x_{iv} = \begin{cases} 0 & \text{for an incorrect response} \\ 1 & \text{for a correct response} \\ \text{NA} & \text{for a missing response.} \end{cases} \quad (1)$$

The missing propensity, ξ, is modeled on the basis of the missing data indicators d_{iv}. The missing data indicators are defined as

[2]Certainly, results from simulation studies that compare different missing data approaches depend on how missing values are induced. Note, however, that the studies we mention used various methods to generate missing data, all arriving at similar conclusions.

$$d_{iv} = \begin{cases} 0 & \text{if } x_{iv} \text{ was not observed} \\ 1 & \text{if } x_{iv} \text{ was observed,} \end{cases} \quad (2)$$

so that for each x_{iv} = NA, $d_{iv} = 0$. Note that higher values on the missing propensity indicate fewer missing responses.[3] The probability that person *v* gives a correct response to item *i* is described as a function of person ability, θ_v, and item difficulty, β_i,

$$p(x_{iv} = 1|\theta_v, \beta_i) = \frac{\exp(\theta_v - \beta_i)}{1+\exp(\theta_v - \beta_i)}. \quad (3)$$

The probability of *observing* an answer from person *v* on item *i* is modeled as a function of the person's missing propensity, ξ_v, and the threshold parameter describing the difficulty of actually giving an answer—regardless of its correctness—to item *i,* δ_i,

$$p(d_{iv} = 1|\xi_v, \delta_i) = \frac{\exp(\xi_v - \delta_i)}{1+\exp(\xi_v - \delta_i)}. \quad (4)$$

The likelihood for the two-dimensional IRT model can be expressed as

$$L = \prod_{v=1}^{V} \prod_{i=1}^{I} p(x_{iv}|\theta_v, \beta_i) p(d_{iv}|\xi_v, \delta_i). \quad (5)$$

[3]The term *missing propensity* is ambiguous, as higher values on this variable indicate fewer missing values. It would be more intuitive to code the missing data indicators reversely or to label ξ the *response propensity*. However, we adhere to the existing literature, where the term *missing propensity* and the respective coding have been established (see, e.g., Holman & Glas, 2005; Rose et al., 2010).

The joint distribution between θ and ξ is considered in the measurement model, thus taking nonignorable missing values into account (Holman & Glas, 2005; O'Muircheartaigh & Moustaki, 1999). Several studies have demonstrated that model-based approaches perform well in retrieving adequate item and person parameter estimates, even when missing values are MNAR (Holman & Glas, 2005; O'Muircheartaigh & Moustaki, 1999; Pohl et al., 2014; Rose, 2013; Rose et al., 2010).

So far, few studies have investigated the influence of missing data approaches on parameters of the structural model.[4] Considering that researchers using data from low-stakes large-scale assessment studies are typically most concerned with parameters relating competence to explanatory variables (e.g., gender, SES), surprisingly little attention has been paid to the influence of missing values on parameters of the structural model. Bias on parameters of the measurement model is not directly transferable to parameters of the structural model. For example, if item and person parameters are equally biased at all levels of the explanatory variable, the estimated relationship coefficient (e.g., between competence and the explanatory variable) might not be strongly affected. If variances and covariances are biased due to an inadequate missing data approach, however, relationship coefficients can be severely affected. So far, studies have shown rather inconsistent results regarding different missing data approaches and estimated latent ability variances, reporting overestimation, underestimation, or no bias at all (Custer et al., 2012; Rose, 2013; Rose et al., 2010). Regarding explanatory variables and interest in their relationship with competence, it is also noteworthy that omitted and not-reached items

[4]We borrow this term from Structural Equation Modeling (SEM) in order to clearly distinguish between the structural model and the measurement model. The structural model is the part of the model specifying correlational links between the latent variables and explanatory variables. The measurement model relates the manifest items to the respective latent variables.

typically relate to potential explanatory variables such as gender, ethnicity, and other competences (Köhler, Pohl, & Carstensen, 2015; Koretz, Lewis, Skewes-Cox, & Burstein, 1993). Glas, Pimentel, and Lamers (2015) investigated the impact of explanatory variables that correlate with missing propensity and competence on the estimation of item difficulty parameters. They found that item parameter estimation was biased when the explanatory variables were not considered in the structural model. The authors thus demonstrated the impact of unaccounted explanatory variables on item parameter estimation. Brown, Svetina, and Dai (2014) investigated whether parameters of the structural model are affected. They used data from NAEP to explore whether an estimated competence difference between various subgroups changes depending on the missing data approach employed. They found virtually no discrepancies in the estimated competence differences when applying four missing data approaches: *incorrect scoring*, *ignoring* missing values, *mean substitution*, and *multiple imputation*. However, the amount of missing values in their data was proportionally small (2%), and model-based approaches were not evaluated.

In our study, we investigated the influence of different missing data treatments on parameters of the structural model considering (a) varying sizes of dependencies between ability, omission propensity, and the explanatory variable, thus generating MCAR, MAR, and MNAR missing data conditions, and (b) varying amounts of missing values. We solely focused on the propensity to omit items, evaluating the performance of (1) the model-based approach as proposed by Holman and Glas (2005; see Equation 5), (2) treating omitted items as incorrect responses, and (3) ignoring missing values. We expected the three investigated missing data approaches to perform differently in recovering parameters of the structural model. Because the model-based approach takes nonignorable missing values into account, we expected this approach to perform best, regardless of (a) and (b). Considering the other two approaches, we

assumed that the amount of nonignorable missing values would impact the bias. We further proposed that the size of the correlation between the explanatory variable and the omission propensity would affect the accuracy of the estimated structural model parameter, because a stronger relationship indicates greater differences in the amount of missing values with respect to the explanatory variable. Since previous studies showed that when treating missing values as incorrect responses, the bias of the item and person parameter estimates was greater than for ignoring missing values, we expected similar results for parameters of the structural model. The amount of missing values should affect the structural parameters in such a way that more missing values would lead to greater bias.

Note that in simulated data, the true parameter values are known, but it is questionable whether the missing data mechanism was induced according to the mechanism in real data. In real data, on the other hand, the missing data mechanism is the mechanism that truly occurred, but the true item and person parameters are unknown. We aimed to remedy this dilemma by investigating both simulated and empirical data. If the results from our simulation study are similar in empirical settings, we have reason to assume that the missing data mechanism we induced in our simulation was adequate and that our results generalize to real data.

Simulation Study

Design

The goal of the simulation study was to generate data sets with missing values that depend on the ability and the explanatory variable to various degrees. To obtain these data sets, we first generated two separate types of data sets: those holding the response indicators x_{iv} and those holding the missing data indicators d_{iv}. We used the Rasch model (Rasch, 1960) to calculate the probabilities for a correct response, $p(x_{iv} = 1)$. The probability depends on person ability, θ_v, and item difficulty, β_i. The probabilities for responding to the item, $p(d_{iv} = 1)$, were generated using

the 2PL model (Birnbaum, 1968) and thus depend on the omission propensity of the person ξ_v, the difficulty of giving an answer to the item δ_i, and the strength of the item to discriminate between people with high and low omission propensity levels, α_i. We chose a 2PL model as our generating model for the missing data indicators because our preliminary analyses showed that it more adequately represents the missing data process found in empirical data. Note that using a 2PL model for the data generation also ensures that our data generating model differs from the model-based approach we later aim to evaluate. Because the true missing data process is usually unknown in empirical data, we aimed to refrain from unjustly preferring any of the methods we investigate by inducing the missing data according to the analyzing model.[5]

In the first step of obtaining x_{iv} and d_{iv}, we generated item parameters β_i, δ_i, and α_i, and person parameters θ_v, ξ_v, and Z_v with $I = 20$ items and $V = 1000$ persons. We used the R package MASS (Venables & Ripley, 2002) in R (R Development Core Team, 2014), drawing the item parameters β_i and δ_i from a bivariate normal distribution with means fixed at 0, variances fixed at 1, and the covariance fixed at .5 (β_{min} = -2.99, β_{max} = 2.02; δ_{min} = -3.32, δ_{max} = 2.72). We chose the standard normal distribution, since item difficulty parameters are often normally distributed in educational studies (see, e.g., Koller, Haberkorn, & Rohm, 2014). When applying model-based approaches to data from large-scale assessments, several of our analyses showed that the item difficulty parameters for answering an item, δ_i, were slightly skewed. Because the skewness was rather small, we decided to simplify our simulation and draw from the normal distribution. The size of the correlation between β_i and δ_i was chosen to be in accordance with

[5]We are aware that the data generating model still comes closest to the model-based approach. As such, our data generation is not completely unrelated to the analyzing models. However, we aimed to establish a data generating model that closely maps the missing data mechanism in empirical data by using generating parameters we found in empirical studies when applying the model-based approach.

the relationship typically found in competence tests of large-scale assessments (see, e.g., Pohl, Haberkorn, Hardt, & Wiegand, 2012; Rose et al., 2010). We specified discrimination parameters, α_i, for the 20 items measuring the omission propensity. We used 20 equally spaced points, covering the range between .5 and 2.5. High values indicate that the item well discriminates between people with many and people with few missing values. We chose such a broad range of discrimination parameters to induce an extreme missing data mechanism. This increases the difference between our data generating model and the model-based approach of the subsequent analyses, in which the item discrimination parameters were set to 1.

The person parameters for the ability level, θ_v, the omission propensity, ξ_v, and the explanatory variable, Z_v, were drawn from a multivariate normal distribution, using the R package mvtnorm (Genz et al., 2014). In our study, we focus on investigating a continuous explanatory variable, since many educational studies address continuous variables such as SES, amount of time studied, personality traits, and attitudes (see, e.g., Blossfeld, Roßbach, & von Maurice, 2011; OECD, 2012). To vary the amount of missing data in the item responses, the mean of the omission propensity variable, ξ, was varied. It was fixed at 0 and 2.5 in two respective conditions. Fixing it at 0 resulted in approximately 50% missing data, which we chose in order to enhance the effects. Changing the mean to 2.5 resulted in a realistic amount of approximately 10% missing data (see, e.g., Cosgrove & Cartwright, 2014; Koretz et al., 1993; OECD, 2012). The means of θ and Z were set to 0 in both conditions. The variances of all three variables, ξ, θ, and Z, were fixed at 1. Besides varying the amount of missing values in the data, we also varied the size of the correlation between the omission propensity and *Z*, with $r(\xi_v, Z_v)$ = 0, .1, .3, and .5, thus considering the conditions of no correlation as well as low, medium, and high correlations according to Cohen (1988). Furthermore, the size of the correlation between ability and omission propensity given *Z* was varied, so that $r(\theta_v, \xi_v \mid Z_v)$ = 0, .2, .4, and .6. These

values were chosen in order to cover a rather wide range of possible correlations. We varied the correlations in order to obtain data sets with different missing data conditions (see Table 1). For the data sets that were generated under the condition of $r(\theta_v, \xi_v \mid Z_v) = 0$ and $r(\xi_v, Z_v) = 0$, the missing data are MCAR; they depend on neither observed nor unobserved variables. When $r(\theta_v, \xi_v \mid Z_v) = 0$ and $r(\xi_v, Z_v) > 0$, the missing data are MAR, since they depend on the observed variable *Z*, but they are conditionally independent of any unobserved variables, given *Z*. For all generated data sets in which $r(\theta_v, \xi_v \mid Z_v) > 0$, persons who would not have answered the item correctly were also more likely to omit the item. The missing values are thus MNAR.

Table 1

Simulation Conditions for Generating Ability, θ_v, Missing Propensity, ξ_v, and Explanatory Variable, Z_v, and the Resulting Missing Data Mechanism Induced in the Simulated Data

$r(\theta_v, Z_v)$	$r(\theta_v, \xi_v)$	$r(\xi_v, Z_v)$	$r(\theta_v, \xi_v \mid Z_v)$	Missing mechanism
0.20	0.00	0.00	0.00	MCAR
0.20	0.02	0.10	0.00	MAR
0.20	0.06	0.30	0.00	MAR
0.20	0.10	0.50	0.00	MAR
0.20	0.20	0.00	0.20	MNAR
0.20	0.21	0.10	0.20	MNAR
0.20	0.25	0.30	0.20	MNAR
0.20	0.27	0.50	0.20	MNAR
0.20	0.39	0.00	0.40	MNAR
0.20	0.41	0.10	0.40	MNAR
0.20	0.43	0.30	0.40	MNAR
0.20	0.44	0.50	0.40	MNAR
0.20	0.59	0.00	0.60	MNAR
0.20	0.60	0.10	0.60	MNAR
0.20	0.62	0.30	0.60	MNAR
0.20	0.61	0.50	0.60	MNAR

The parameters of the covariance matrix when drawing parameters θ_v, ξ_v, and Z_v for each *v* from the multivariate normal are displayed in Table 1. Note that we fixed the values of the partial correlation between θ_v and ξ_v (see Column 4), which naturally resulted in altering bivariate correlations between θ_v and ξ_v (see Column 2). The correlation between ability and *Z* was fixed at $r(\theta_v, Z_v) = .2$ for all data sets. Note that $r(\theta_v, Z_v) = .2$ was the coefficient we later aimed to retrieve. Altogether, three factors were varied, resulting in a 4 (correlation between the ability and the omission propensity) x 4 (correlation between the omission propensity and *Z*) x 2 (amount of missing data) design with = 32 cells. The number of replications for each of the possible combinations was $w = 100$.

In the second step toward obtaining x_{iv} and d_{iv}, we calculated the probabilities for a correct response, $p(x_{iv} = 1)$, and the probabilities for giving a response to the item, $p(d_{iv} = 1)$. $p(x_{iv} = 1)$ was calculated under the Rasch model, using the previously generated person and item parameters θ_v and β_i. $p(d_{iv} = 1)$ was calculated under the 2PL model, using ξ_v, δ_i, and α_i. We compared the probabilities for a correct response, $p(x_{iv} = 1)$, to values randomly drawn from a uniform distribution on the interval (0, 1). When $p(x_{iv} = 1)$ exceeded the randomly drawn value, the response indicators were scored $x_{iv} = 1$, and $x_{iv} = 0$ otherwise. The same was done with regard to the probabilities for giving a response, $p(d_{iv} = 1)$, thus obtaining d_{iv}. The response indicators m_{iv} for the data sets containing missing values were derived from x_{iv} and d_{iv} and were defined as

$$m_{iv} = \begin{cases} 0 & \text{for } x_{iv} = 0 \text{ and } d_{iv} = 1 \\ 1 & \text{for } x_{iv} = 1 \text{ and } d_{iv} = 1 \\ \text{NA} & \text{for } d_{iv} = 0. \end{cases} \quad (6)$$

We subsequently used latent regression models to estimate the relationship between the latent ability variable θ and the manifest variable Z. We first analyzed the data sets without

missing values with a unidimensional latent regression model in order to establish a frame of reference and to illustrate the relationship between θ and Z in the complete data. For data sets containing missing values**,** we considered three different missing data approaches: (1) apply the model-based approach and include the latent omission propensity ξ_v in the measurement model, (2) ignore missing values in the estimation, and (3) treat missing values as incorrect responses. All three approaches were applied to 3200 data sets. The first approach (1) was based on the between-item multidimensional IRT model by Holman and Glas (2005; see Equation 3). We included Z as a predictor for both the ability and the omission propensity, resulting in a two-dimensional latent regression model. The marginal maximum likelihood (MML) estimation equation of the multidimensional latent regression model can be expressed as

$$L = \prod_{v=1}^{V} \int \int \prod_{i=1}^{I} p(x_{iv}|\theta_v, \beta_i) p(d_{iv}|\xi_v, \delta_i) g(\theta_v, \xi_v|Z_v, \eta, \sum) \, d\theta_v \, d\xi_v, \qquad (6)$$

where $g(\theta_v, \xi_v|Z_v, \eta, \sum)$ represents the density of the conditional common distribution of θ_v and ξ_v given Z, which is assumed to be bivariate normal. Z_v is the value of person v on the variable Z, η are the regression coefficients, and $\sum$ represents the covariance matrix of the residuals. Note that all item discrimination parameters are set to 1 and, thus, do not appear in the equation. For the second (2) and third (3) approaches, unidimensional latent regression IRT models were estimated. Equation 6 thus simplifies to

$$L = \prod_{v=1}^{V} \int \prod_{i=1}^{I} p(x_{iv}|\theta_v, \beta_i) g(\theta_v|Z_v, \eta, \sigma^2) \, d\theta_v, \qquad (7)$$

where σ^2 is the residual variance of θ_v. When treating missing items as incorrect answers (approach 3), each m_{iv} = NA was replaced by m_{iv} = 0. All models were estimated using the R package TAM (Kiefer, Robitzsch, & Wu, 2014). To compute the integrals, we used Gauss-Hermite quadrature with 20 nodes per dimension. A minimum deviance change of .0001 was chosen as the convergence criterion.

In a last step, we calculated the mean standardized regression coefficient across all 100 replications for each of the considered combinations. For all analyses, we constrained the intercept to be 0. We used the directly estimated regression coefficient from the model to estimate the mean standardized regression coefficient, and we divided it by the unconditional variance of the latent ability (see Wu, Adams, Wilson, & Haldane, 2007). Note that the latent regression model in Equation 7 closely relates to the estimation method commonly used in large-scale assessments (see, e.g., Adams & Wu, 2000; von Davier, Sinharay, Oranje, & Beaton, 2006). Our model is less complex, since we only consider one competence dimension, one conditioning variable *Z*, and only one group of examinees. To establish whether the deviation between the estimated and true regression coefficients is actually meaningful, we interpret the standardized regression coefficient in terms of an effect size. Many researchers apply Cohen's rules of thumb (Cohen, 1988) when drawing inferences from effect sizes, so we consider changes above .1—a small effect size—to be practically relevant.

Results

The mean standardized regression coefficients from the latent regression of ability on *Z* over the 100 replications are depicted in Figure 1. As is evident from the figure, the mean standardized regression coefficients were close to the generating parameter ($r(\theta_v, Z_v)$ = .2, see grey line) in all conditions for the complete data analyses. The two slight deviations from the grey line (top row, left picture: $r(\theta_v, \xi_v \mid Z_v)$ = 0, $r(\xi_v, Z_v)$ = 0, 10% condition; top row, third

picture: $r(\theta_v, \xi_v \mid Z_v) = .4$, $r(\xi_v, Z_v) = .3$, 10% condition) are due to relatively few replications, $w = 100$.

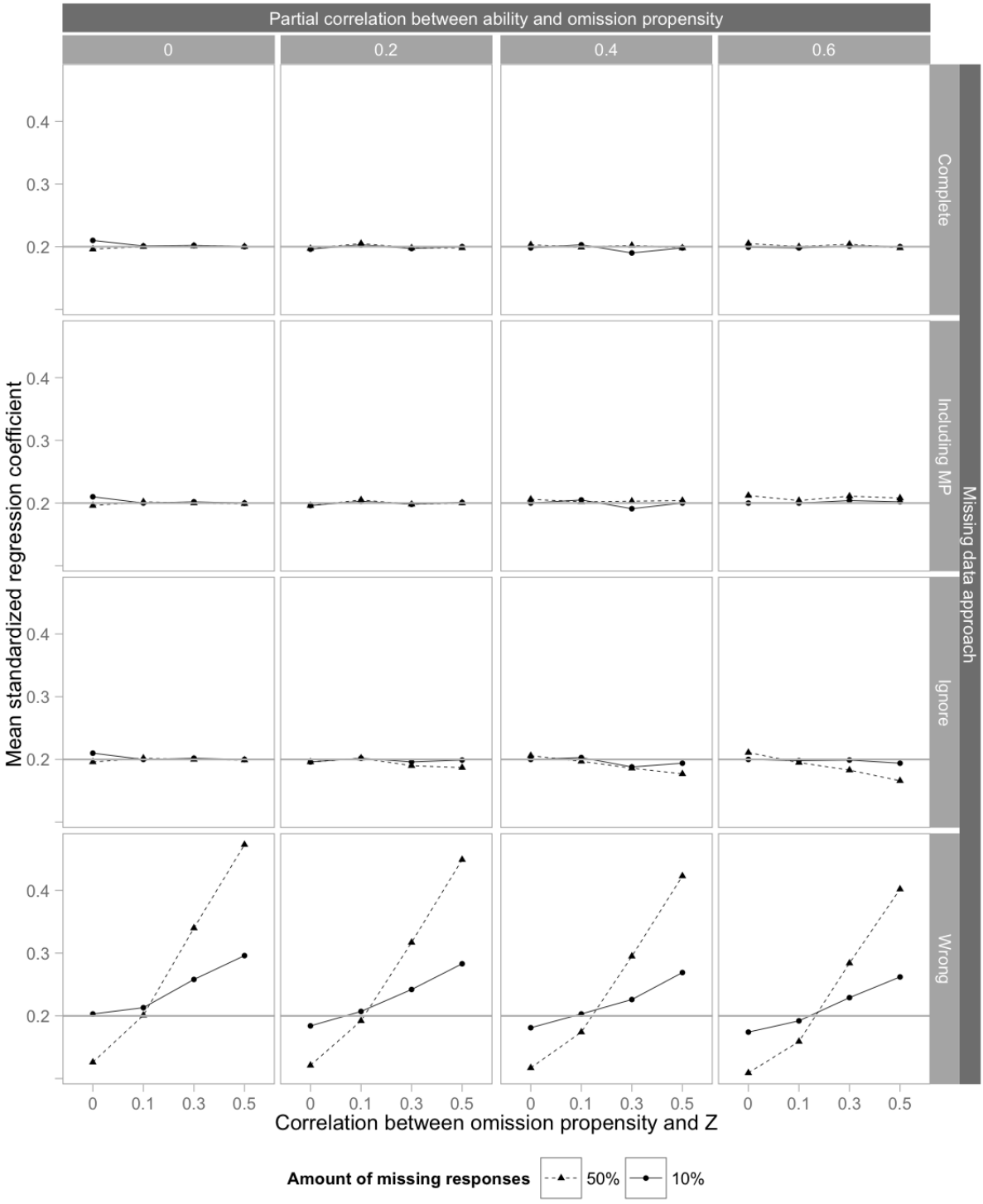

Figure 1. Estimated mean standardized regression coefficients of ability on the variable *Z* across 100 replications for all realized combinations. MP = missing propensity.

Inclusion of the missing propensity.

In line with our expectations, the model-based approach succeeded in retrieving unbiased regression coefficients. Nearly identical results were obtained in the complete case analysis. The estimates are slightly higher only in the last column, in which the omission propensity greatly depends on ability and the amount of missing values was 50%. This difference is hardly noticeable, however, and can thus be neglected. Overall, the model-based approach is robust against the misspecification of item discrimination parameters, and it accurately estimates structural parameters of the model.

Ignoring missing values.

As is evident from the figure, ignoring missing values resulted in unbiased estimates of the regression coefficient when the missing values were either MCAR or MAR. Furthermore, the mean standardized regression coefficient was close to the generating parameter in the conditions where the partial correlation between omission propensity and ability was $r(\theta_v, \xi_v \mid Z_v) = 0$ and $r(\theta_v, \xi_v \mid Z_v) = .2$. In the conditions where $r(\theta_v, \xi_v \mid Z_v) = .4$ and $r(\theta_v, \xi_v \mid Z_v) = .6$, the regression coefficients were unbiased for low correlations between the omission propensity and Z.[6] For higher correlations between the omission propensity and Z, the regression coefficient was slightly underestimated in the 50% missing condition.[7]

[6] In the condition with $r(\theta_v, \xi_v \mid Z_v) = .6$, $r(\xi_v, Z_v) = 0$ and 50% missing values, the regression coefficient seems somewhat overestimated. However, the same is true for the model including the missing propensity and the model using the complete data set and the missing propensity (the latter is not shown in the figure). Therefore, the overestimation is not due to the missing data treatment and can thus be neglected.

[7] In the condition with $r(\theta_v, \xi_v \mid Z_v) = .4$, $r(\xi_v, Z_v) = .3$ and 10% missing values, the regression coefficient seems somewhat underestimated. However, the same is true for the complete case analysis and the model including the missing propensity. The low estimate most likely resulted from a random error due to only $w = 100$ replications.

To explain this underestimation, we estimated EAP person parameter estimates and compared them with the true values at different levels of Z.[8] The underestimation resulted from bias in ability estimates that varied at different levels of Z: The ability estimates were, on average, overestimated for people with lower Z scores and were, on average, underestimated for people with higher Z scores. This was due to the underspecification of the model, as ξ was not included in the measurement model. Note, however, that the bias is rather small in the estimated regression coefficient when ignoring missing values. In all conditions, the difference between the true and the estimated regression coefficient was less than .034. From a practical point of view, ignoring missing values rendered similar results to using the model-based approach.

An interesting aspect that is also evident in our results concerns the discrepancy between bias of parameters of the measurement model and bias of parameters of the structural model. In our study, individual ability estimates were biased in the condition of $r(\theta_v, \xi_v \mid Z_v) = .6$ and $r(\xi_v, Z_v) = 0$. Persons at the lower end of the ability distribution range especially profited from ignoring missing responses. The estimated regression coefficient, however, showed no bias, and it adequately reflected the dependency between ability and the explanatory variable. This illustrates that bias on the individual level does not necessarily affect the structural model.

Scoring missing values as incorrect.

When treating missing values as incorrect, the estimated regression coefficients were substantially biased. In most conditions with low correlations between the omission propensity and Z, the regression coefficients were underestimated. In the conditions where the correlation between the omission propensity and Z was greater than .1, the regression coefficients were

[8]The individual ability estimates were calculated with the explanatory variable in the conditioning model.

overestimated (see Figure 1). The bias occurred regardless of the relationship between ability and missing propensity.

To explain these results, we again investigated bias on the individual level at different levels of *Z*. Individual ability values were, on average, underestimated under incorrect scoring. This general underestimation resulted from imputing each missing value with a score of 0. Naturally, imputing an incorrect answer for an unobserved correct answer yields a lower ability score (compared to the ability that would be scored if the item had been observed). In our simulation, people with high ability levels showing more correct answers in the complete data set than people with low ability levels were more severely underestimated compared to people with low ability levels. Furthermore, the average bias in ability estimates varied for different levels of *Z*. In the data sets where missing values did not depend on *Z*, persons had an equal amount of missing values, irrespective of *Z*. However, since the generating parameter for the correlation between ability and *Z* was $r(\theta_v, Z_v) = .2$, people with higher *Z* scores had higher ability levels, and their ability was therefore, on average, more underestimated than the ability of people with lower *Z* scores. Thus, the standardized regression coefficient when regressing ability on *Z* was underestimated. In the data sets where missing values greatly depended on *Z*, the ability estimates were, on average, more underestimated for people with low *Z* scores than for people with high *Z* scores, since the latter obtained substantially fewer missing values. As a result, the slope of the standardized regression coefficient when regressing ability on *Z* was substantially overestimated.

The size of the partial correlation between the omission propensity and the ability also influenced the estimated regression coefficient when treating missing values as incorrect. For a constant correlation between the omission propensity and *Z*, the estimated regression coefficient decreased as the partial correlation between ability and the omission propensity increased. This was mainly linked to a greater estimated total variance of ability in the conditions where ability

depended on both *Z* and the omission propensity. The proportion of variance *Z* explained in ability was therefore comparably smaller, thus reducing the standardized regression coefficient.

The impact of treating missing values as incorrect responses on the estimated regression coefficient was quite severe. In eight of the 32 conditions, the difference between the true and the estimated regression coefficient exceeded .1. Here, the conclusions a researcher would draw from the regression analyses differ substantially from the conclusions one might draw when applying the model-based approach or when ignoring missing values. Note that seven of the eight conditions showing severe bias contained 50% missing values, and only one contained 10% missing values. This means that in the condition with 10% missing values, incorrect scoring did not lead to extreme differences in the conclusions drawn for the relationship between ability and *Z*. Consider, however, that for substantive analyses, usually, more than one explanatory variable is used, and bias would thus no longer be limited to only one parameter of the structural model. In some PISA countries, however, the omission rate reaches 20%, not including not-reached items (which also need to be considered). The bias due to incorrect scoring we found in our simulation study should therefore not be discounted.

In sum, the results indicate that for data in which the missing process is similar to our simulated data, accurate parameters are retrieved when applying the model-based approach and including the omission propensity in the measurement model, even if the model for the omission propensity imprecisely represents the true missing data process. When ignoring missing values, bias is very small and only present at high correlations between ability, missing propensity, and *Z*. The size of the bias is hardly substantial, even for high amounts of missing data. For incorrect scoring, the bias in parameter estimates of the structural model can be quite substantial, especially given a high amount of missing data.

Empirical Examples

Method

Finally, we applied the different missing data approaches to empirical data in order to investigate whether our results were transferable to real competence test data and to examine how the coefficients from group analyses are affected when the missing values relate to ability and the explanatory variable. Overall, we conducted four empirical studies, using data from two different large-scale assessments: PISA 2009 and NEPS. Examples 1 and 2 comprise the PISA 2009 data on reading literacy in Albania, differing with respect to the explanatory variable. In Example 1, we investigated the relationship between reading competence and a continuous explanatory variable. In Example 2, we investigated the relationship between reading competence and a bivariate explanatory variable. Example 3 comprised a subsample of the data from Examples 1 and 2, namely, only students who received test booklet 10. Example 4 involves data from the mathematics domain in the NEPS adult cohort. We chose these examples to probe whether the results from our simulation are generalizable across various empirical settings. Compared to examples 1 and 3, a bivariate variable is used in examples 2 and 4. In examples 3 and 4, all examinees responded to all items, while a rotated test design was used in examples 1 and 2.[9] The NEPS data have fewer omitted items (4%) than the PISA 2009 Albania data (16% overall, 21% in subsample test booklet 10) and thus serve to illustrate that the effect of different missing data treatments diminishes as the amount of missing values decreases.

The explanatory variables in the four examples were (1) reading enjoyment; (2) one item from the reading attitude scale ("I only read if I have to"), where we collapsed the categories *strongly disagree* and *disagree* and the categories *agree* and *strongly agree*; (3) reading

[9]Rotated test designs produce missing responses by design, which are considered MCAR. They should not affect parameter estimation (Rubin, 1976); we therefore refrained from simulating this aspect in our study. Nevertheless, we wanted to illustrate that our results from the simulation study hold in empirical settings with various test designs.

enjoyment; (4) and gender. The number of items in the examples was (1) 90, (2) 90, (3) 12, and (4) 21. The number of examinees was (1) $V = 3{,}779$, (2) $V = 3{,}732$, (3) $V = 339$, and (4) $V = 2{,}333$. Note that since we only focused on omitted items in our simulation study, examinees with not-reached items were excluded from the analyses.[10] Students with missing values on the explanatory variable were also excluded, which explains the different sample sizes in (1) and (2).

Before investigating the relationships between the explanatory variable and ability, we used unidimensional Rasch models for reading ability and omission propensity, respectively, examining relevant parameters such as item difficulties, their respective standard errors, the correlation between the item parameters of the two models, and the variances of the latent variables. These parameters inform about differences and similarities between our simulation study and the real data examples. Subsequently, we applied the three different missing data approaches to estimate the standardized regression coefficient of ability on *Z*: (1) including the omission propensity in the measurement model, (2) ignoring missing values, and (3) treating missing values as incorrect responses. As before, we used the regression coefficient directly estimated from the model and divided it by the unconditional ability variance in order to obtain the standardized regression coefficient. The intercept was constrained to be 0. In the real data, we expected to find effects similar to our simulation study.

Results

Relevant model parameters that inform about the similarities between the real data examples and our simulation study are illustrated in Table 2.

[10]Note that not-reached and omitted items are treated differently in the scaling of PISA: Not-reached items are ignored for the international calibration of item difficulties and are treated as incorrect for the student score generation; omitted items are treated as incorrect responses at all stages (Adams & Wu, 2002; Martin et al., 2000).

Table 2

Missing values and model parameters of the empirical data examples.

	Data		
	PISA 2009 Albania[a]	PISA 2009 Albania Test booklet 10	NEPS adult cohort, mathematics domain
Range Omissions per I	21-668	20-154	0-24
β_{min} (*SE*), β_{max} (*SE*)	-3.72(.16), 3.56(.18)	-1.77(.17), 2.20(.22)	-3.38(.09), 1.55(.06)
$\bar{\beta}_i$	-0.23	0.26	-0.49
δ_{min} (*SE*), δ_{max} (*SE*)	-5.38(.21), 0.68(.08)	-4.37(.27), -0.36(.15)	-7.00(.32), -0.75(.05)
$\bar{\delta}_i$	-2.90	-2.46	-4.21
$r(\beta_i, \delta_i)$	.53	.54	.67
$r(\theta_v, \xi_v \mid Z_v)$	.60	.35	.49
$r(\xi_v, Z_v)$	.10	.32	.31
σ_θ^2, σ_ξ^2	0.99, 4.15	0.77, 5.22	1.84, 3.53

Note. [a]Nearly identical data sets were used for empirical Examples 1 and 2. In this table, results from Example 2 are shown. Except for the correlation between the explanatory variable and *Z*, which was $r(\xi_v, Z_v)$ = .30 in Example 1, the parameters from both examples hardly deviated.

Note that in our examples, the distribution of the item difficulties for the missing propensity δ_i is rather skewed and therefore deviates from the setup of our simulation study, where we drew the item parameters from a standard normal. The correlation between the difficulty of answering an item correctly and the difficulty of giving an answer to the respective item in the PISA 2009 data was close to the generating parameter we used in our simulation study ($r(\beta_i, \delta_i)$ = .5) and even higher in the NEPS data. The conditional correlation between ability and missing propensity was high in examples 1, 2, and 4 and medium in the Albania subsample. Thus, they closely mirror our simulation study conditions $r(\theta_v, \xi_v \mid Z_v)$ = .4 and $r(\theta_v, \xi_v \mid Z_v)$ = .6. The correlation between the omission propensity and *Z* was $r(\xi_v, Z_v)$ = .10 in Example 2 and close to $r(\xi_v, Z_v)$ = .30 in Examples 1, 3, and 4, thus mirroring the small and medium correlations of our simulation study.

Table 3 shows the estimated standardized regression coefficients, standardized standard errors, and the respective confidence intervals. The results illustrate that the estimated relationship between ability and reading enjoyment varies depending on the applied missing data approach.

Table 3

Regressing ability on (1) reading enjoyment, (2) reading attitude (1 item, dichotomously scored), (3) reading enjoyment, and (4) gender, using different missing data approaches.

Data	Model	b_{stz}	se_{stz}	95% CI
PISA 2009 Albania[a]	Inclusion of missing propensity	.316	0.012	0.293-0.339
	Ignoring missing values	.301	0.013	0.276-0.327
	Incorrect scoring	.340	0.012	0.316-0.364
PISA 2009 Albania[a]	Inclusion of missing propensity	.123	0.002	0.120-0.127
	Ignoring missing values	.124	0.002	0.119-0.129
	Incorrect scoring	.130	0.002	0.126-0.134
PISA 2009 Albania test booklet 10	Inclusion of missing propensity	.228	0.050	0.129-0.326
	Ignoring missing values	.225	0.052	0.124-0.326
	Incorrect scoring	.338	0.045	0.249-0.427
NEPS adult cohort, mathematics domain	Inclusion of missing propensity	.329	0.015	0.300-0.358
	Ignoring missing values	.327	0.016	0.296-0.357
	Incorrect scoring	.339	0.015	0.309-0.351

Note. b_{stz} = standardized regression coefficient; se_{stz}= standardized standard error of regression coefficient; 95% CI = lower and upper limits of 95% confidence interval.

[a]The data sets slightly differ in sample size, since we deleted examinees with missing values on the predictor variables, and examinees with a missing value on *reading enjoyment* (Example 1) differed from examinees with a missing value on *reading attitude* (Example 2)

Altogether, the real data results mirror the findings from our simulation study, where in the more extreme conditions, the approach of ignoring missing values resulted in similar

regression estimates and the approach of treating missing values as incorrect responses resulted in higher estimates. [11] As is evident from the table, the effect is consistent in all examples, and it diminishes as the amount of missing values in the data decreases. In the PISA 2009 Albania test booklet 10 example, the difference between the estimated regression coefficients even exceeded .1, meaning that the conclusion a researcher would draw from the regression analysis when missing values are scored as incorrect values substantially differs from the conclusion when either of the other approaches is employed.

Discussion

The goal of our study was to evaluate different approaches to treating missing item responses in recovering parameters of the structural model, that is, the model specifying relationships between latent variables and explanatory variables such as gender or SES. Missing values often relate to these explanatory variables as well as the assessed competence, and we aimed to investigate how the missing data treatment affects the parameter estimates of the structural model given different relations between missing values, an explanatory variable, and competence. We investigated the approaches of ignoring missing values and treating them as incorrect, which are more frequently used in the scaling of large-scale assessments, as well as a recently developed model-based approach.

The three approaches perform differently depending on the present missing data mechanism. The main finding was that ignoring missing values and the model-based approach

[11] To give an idea of what the difference can imply with regard to individual ability parameters, we used Example 1 and estimated WLE estimates for the two examinees who scored lowest and highest on the enjoyment variable. For the two examinees, the difference on the achievement level (in logits) was 2.6 when the missing propensity was included in the scaling model, 2.57 when missing values were ignored, and 3.12 when missing values were treated as incorrect values.

led to nearly identical parameter estimates. The model-based approach only outperformed the approach of ignoring missing values when the amount of missing values in the data was large and when the probability for a missing value greatly depended on the latent ability variable (after controlling for the explanatory variable) and on the explanatory variable. Incorrect answer substitution resulted in considerably different estimates of the structural model when the amount of missing values was large. The relationship between competence and the explanatory variable was either over- or underestimated, regardless of whether missing values were MCAR, MAR, or MNAR.

Our results confirm and enhance previous research. In line with findings regarding bias on person and item parameters, the approach of ignoring missing values leads to accurate estimates of regression coefficients when missing values are MAR. When missing values are MNAR, the parameter estimates of the structural model are hardly affected. At first glance, this might seem to contradict findings from previous studies that showed that item and person parameters are biased when missing values are nonignorable. If the missing propensity and the explanatory variable are uncorrelated, however, the estimated relationship between competence and the explanatory variable remains unaffected. As the dependency between the missing propensity and the explanatory variable increases, bias on the individual level varies for different levels of the explanatory variable and thus distorts the relationship between the explanatory variable and competence. Note, however, that bias was actually rather small and only found for high amounts of nonignorable missing data. In the more realistic condition with only 10% missing values, the parameter estimates of the structural model were extremely robust against violations of nonignorability. Overall, our study showed that while item and person parameter estimates are biased in the condition of nonignorable missing values, structural parameters are only affected

when the missing propensity is closely related to the explanatory variable as well. Even under these conditions, bias is limited and only appears when the amount of missing values is high.

In line with our expectations, parameters of the structural model were more significantly biased when missing values were substituted with incorrect answers. A new finding from our study is that the relationship between competence and an explanatory variable was biased in simulated conditions where the missing values were MAR or even MCAR. If our simulated missing data mechanism comes close to the missing data mechanism of actual data, our results stress the importance of rethinking using the incorrect scoring method. Although substantial deviations mostly occurred for high amounts of missing data—a situation hardly encountered in large-scale assessments—the parameters will potentially be more biased using incorrect scoring. Cases including several explanatory variables and not-reached items might even enhance the effects we found.

The setup of our simulation study differed from real large-scale assessment data in several aspects. It only comprised a limited number of items and persons, and all persons responded to all items. Thus, the simulated data did not match the multi-matrix sampling design that is typical for many operational tests. Also, no booklet design was included. How results from the simulation study generalize to longer tests with more examinees and more complex test designs needs to be thoroughly investigated in future studies. The empirical examples indicate that the effects from the simulation study might be transferable to empirical settings. In the empirical examples, different test designs were used, including multi-matrix sampling and booklet designs. For all test designs, the results from the empirical analyses closely resembled the results from the data that were simulated under less complex test designs but with similar amounts of missing data and similar relationships between the missing propensity, the explanatory variable, and competence. This is a first indication that the results from our simulation study may be generalizable to more

complex test settings such as those used in large-scale assessment studies. In order to draw conclusions about the generalizability of our results to other test settings, further research is necessary.

Although our data-generating model involved a continuous explanatory variable, the results from the empirical examples already indicate that they also generalize to binary variables. To further illustrate this, we simulated an example with a binary variable that mirrored the most-extreme conditions of our simulation study: $r(\theta_v, \xi_v \mid Z_v) = .6$, $r(\xi_v, Z_v) = .5$, 50% missing values. Again, we analyzed the relationship between the explanatory variable and ability using the three missing data approaches. Our results were the same: Applying the model-based approach came closest to retrieving the true regression coefficient, ignoring missing values led to a slight underestimation of the regression coefficient, and treating missing values as incorrect responses resulted in a largely overestimated regression coefficient (see Table 1 of the Appendix). For ordinal data, generalization of our results is less straightforward. In a regression with an ordinal variable as the explanatory variable, the categories *k* of the variable need dummy coding so that each category is compared to a predetermined reference category (e.g., the null-category). This results in *k*-1 regression coefficients. Each of the regression coefficients could potentially be biased due to the missing data treatment. The amount of bias would depend on the amount of missing values in the respective groups, the competence differences between the groups, and the dependency between the missing values and competence. Based on our results, we would expect that the more people differ in their amount of missing values and their ability, and the higher the dependency between missing values and ability is, the greater the bias of the regression coefficient for the respective groups will be. Group size, that is, the number of people in each category, might also affect the accuracy of the regression coefficients, especially for unbalanced groups. For example, if few people agreed with the highest category, and this group

of people had the highest number of omitted items, the estimated ability variance might be more biased in this group compared to the reference group. This would in turn affect the estimated regression coefficient. These considerations need to be tested in future studies.

For high-stakes assessments, ignoring missing values should not be the method of choice, since examinees who are aware of the scoring method might simply omit the questions they are unsure of. This would increase the missing data rates considerably. The missing data mechanism in high-stakes assessments should be investigated in future studies. In low-stakes assessments, examinees are typically less motivated to increase their test scores (Jakwerth, Stancavage, & Reed, 1999; Sundre, 1999). Our results also demonstrate that even when the missing propensity highly depends on competencies, the ability parameter estimates remain relatively unbiased. For very high correlations between missing propensity and ability, researchers should rather consider the model-based approach and include the missing propensity in the scaling model.

There are some limitations to our study and the inferences that can be drawn from it. In all simulation studies, data were generated according to certain models. Of course, missing values could be simulated alternatively. A missing value could, for example, be induced depending on whether the response to the item was correct or incorrect (Robitzsch, 2016). Different data generating models may lead to different conclusions about the missing data approach. Since some of the reasons for missing values include inability to answer the item correctly as well as other personality states and traits (cf. Jakwerth et al., 1999; Köhler et al., 2015; Koretz et al., 1993; Pohl et al., 2014; Rose et al., 2010), we generated the missing data such that there would be dependence on ability and an explanatory variable in order to examine whether disregard of these relationships affects the parameter estimates of the structural model. A simulation study cannot mirror all properties of the missing data process in actual empirical data. We tried to cover

various missing data scenarios and missing data mechanisms, and we used empirical examples to validate our conclusions.

We only considered the case where the missing data depend on competence and on one explanatory variable. In empirical data, however, the missing propensity may depend on several other person characteristics (Jakwerth et al., 1999; Köhler et al., 2015; Koretz et al., 1993). It would be interesting to assess how estimates of the structural model are affected when these relationships are accounted for in the data generation and analysis. Furthermore, we only considered missing values due to omitted items. The propensity to not reach items typically differs from the propensity to omit items, and it should be handled separately in the scaling (Moustaki & O'Muircheartaigh, 2000; Pohl et al., 2014; Rose, 2013). It is worthwhile to consider not-reached and omitted items simultaneously and to investigate how the different missing data processes influence parameter estimates of the structural model.

References

Adams, R., & Wu, M. (2002). *PISA 2000 technical report.* Paris, France: Organisation for Economic Co-operation and Development.

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 technical report* (NCES Publication No. 2001-509). Washington, DC: National Center for Education Statistics.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (eds.) (2011). Education as a lifelong process – the German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft: Special Issue 14*.

Brown, N. J. S., Svetina, D., & Dai, S. (2014). *Impact of methods of scoring omitted responses on achievement gaps.* Presentation at the annual meeting of the National Council on Measurement in Education, Philadelphia, PA. Retrieved from http://ceep.indiana.edu/ImplicationsFromNAEP/Brown%202014%20NCME%20Scoring%20omitted%20responses.pdf

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edition). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cosgrove, J., & Cartwright, F. (2014). Changes in achievement on PISA: the case of Ireland and implications for international assessment practice. *Large-scale Assessments in Education, 2*, 1-17. doi:10.1186/2196-0739-2-2

Culbertson, M. (2011, April). *Is it wrong? Handling missing responses in IRT.* Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, LA.

Custer, M., Sharairi, S., & Swift, D. (2012, April). *A comparison of scoring options for omitted and not-reached items through the recovery of IRT parameters when utilizing the Rasch model and Joint Maximum Likelihood Estimation.* Paper presented at the annual meeting of the National Council of Measurement in Education, Vancouver, British Columbia. Retrieved from http://files.eric.ed.gov/fulltext/ED531171.pdf

DeAyala, R. J., Plake, B. S., & Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement, 38*, 213-234. doi:10.1111/j.1745-3984.2001.tb01124.x

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement, 45*, 225-245. doi:10.1111/j.1745-3984.2008.00062.x

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2014). mvtnorm: Multivariate normal and t distributions. R package version 1.0-2. Retrieved from http://CRAN.R-project.org/package=mvtnorm

Glas, C. A. W., & Pimentel, J. L. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement, 68*, 907-922. doi:10.1177/0013164408315262

Glas, C. A. W., Pimentel, J. L., & Lamers, S. M. A. (2015). Nonignorable data in IRT models: Polytomous responses and response propensity models with covariates. *Psychological Test and Assessment Modeling, 57*, 523-541.

Hohensinn, C., & Kubinger, K. D. (2011). Applying item response theory methods to examine the impact of different response formats. *Educational and Psychological Measurement, 71*, 732-746. doi:10.1177/0013164410390032

Holman, R., & Glas, C. A. W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology, 58*, 1-17. doi:10.1111/j.2044-8317.2005.tb00312.x

Jakwerth, P. M., Stancavage, F. B., & Reed, E. D. (1999). *An investigation of why students do not respond to questions* (NAEP Validity Studies, Working Paper Series). Palo Alto, CA: American Institutes for Research. Retrieved from http://www.air.org/sites/default/files/downloads/report/Jakwerth_report_0.pdf

Kastberg, D., Roey, S., Lemanski, N., Chan, J. Y., & Murray, G. (2014). *Technical Report and User Guide for the Program for International Student Assessment (PISA).* (NCES 2014 -025). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Kiefer, T., Robitzsch, A., & Wu, M. (2014). TAM: Test analysis modules. R package version 1.5-2. Retrieved from http://cran.r-project.org/web/packages/TAM/index

Köhler, C., Pohl, S., & Carstensen, C. H. (2014). Taking the missing propensity into account when estimating competence scores—Evaluation of IRT models for non-ignorable omissions. *Educational and Psychological Measurement*. doi: 10.1177/0013164414561785

Köhler, C., Pohl, S., & Carstensen, C. H. (2015). Investigating mechanisms for missing responses in competence tests. *Psychological Test and Assessment Modeling, 57, 499-522.*

Koller, I., Haberkorn, K., & Rohm, T. (2014). *NEPS Technical Report for Reading: Scaling results of Starting Cohort 6 for adults in main study 2012* (NEPS Working Paper No. 48). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Koretz, D., Lewis, E., Skewes-Cox, T., & Burstein, L. (1993). *Omitted and not-reached items in mathematics in the 1990 National Assessment of Educational Progress* (CSE Tech. Rep. No. 357). Los Angeles, CA: Center for Research on Evaluation, Standards and Student Testing.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39,* 247-264.

Martin, M. O., Gregory, K. D., & Stemler, S. E. (2000). *TIMSS 1999 technical report.* Chestnut Hill, MA: Boston College.

Matters, G., & Burnett, P. C. (2003). Psychological predictors of the propensity to omit short response items on a high-stakes achievement test. *Educational and Psychological Measurement, 63*, 239-256. doi: 10.1177/0013164402250988

Mislevy, R. J., & Wu, P.-K. (1988). *Inferring examinee ability when some item responses are missing.* Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (RR 96-30-ONR). Princeton, NJ: Educational Testing Service.

Moustaki, I., & O'Muircheartaigh, C. (2000). A one dimensional latent trait model to infer attitude from nonresponse for nominal data. *STATISTICA*, 259-276.

OECD (2012). *PISA 2009 Technical Report, PISA*, OECD Publishing.

O'Muircheartaigh, C., & Moustaki, I. (1999). Symmetric pattern models: A latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society, Series A: Statistics in Society, 162*, 177-194. doi: 10.1111/1467-985X.00129

Pohl, S., & Carstensen, C. H. (2012). *NEPS technical report: Scaling the data of the competence tests* (NEPS Working Paper No. 14). Bamberg, Germany: Otto-Friedrich-Universität, Nationales Bildungspanel.

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in Item Response Theory models. *Educational and Psychological Measurement, 74,* 423–452. doi: 10.1177/0013164413504926

Pohl, S., Haberkorn, K., Hardt, K., & Wiegand, E. (2012). NEPS Technical Report for Reading – Scaling Results of Starting Cohort 3 in Fifth Grade (NEPS Working Paper No. 15). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

R Development Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Nielsen & Lydiche. (Expanded edition, 1980)

Robitzsch, A. (2016). Zu nichtignorierbaren Konsequenzen des (partiellen) Ignorierens fehlender Item Responses im Large-Scale Assessment. In B. Suchań, C. Wallner-Paschon, & C. Schreiner (Eds.), *PIRLS & TIMSS 2011. Die Kompetenzen in Lesen, Mathematik und Naturwissenschaften am Ende der Volksschule.* Graz: Leykam.

Rose, N. (2013). *Item nonresponses in educational and psychological measurement.* Ph.D. thesis, Friedrich-Schiller-University Jena, Dept. of Methodology and Evaluation Research.

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT)* (ETS Research Report ETS RR-10-11). Princeton, NJ: Educational Testing Service.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581-592. doi:10.1093/biomet/63.3.581

Sundre, D. L. (1999, April). *Does examinee motivation moderate the relationship between test consequences and test performance?* Paper presented at the annual meeting of the American Educational Research Association, Montreal. (ERIC Document Reproduction Service No. ED432588)

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). Statistical procedures used in the National Assessment of Educational Progress (NAEP): Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics.* Amsterdam: Elsevier.

Venables, W. N. & Ripley, B. D. (2002). Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest Version 2.0. Generalised item response modeling software.* Victoria: ACER Press.

Appendix

Table 1

Estimated mean standardized regression coefficients of ability on bivariate explaining variable Z across 100 replications.

Model	b_{stz}
True (generating) parameter	.249
Complete data set	.233
Including missing propensity	.232
Ignoring missing values	.200
Incorrect scoring	.403

Note. b_{stz} = standardized regression coefficient.